

# Superheat: Supervised heatmaps for visualizing complex data

Rebecca L. Barter

Department of Statistics, University of California, Berkeley  
and

Bin Yu

Department of Statistics, University of California, Berkeley

December 7, 2015

## Abstract

Technological advancements of the modern era have enabled the collection of huge amounts of data in science and beyond. Accordingly, computationally intensive statistical and machine learning algorithms are being used to seek answers to increasingly complex questions. Although visualization has the potential to be a powerful aid to the modern information extraction process, visualizing high-dimensional data is an ongoing challenge. In this paper we introduce the supervised heatmap, called superheat, which is a new graph that builds upon existing clustered heatmaps that are widely used in fields such as bioinformatics. Supervised heatmaps have two primary aims: to provide a means of visual extraction of the information contained within high-dimensional datasets, and to provide a visual assessment of the performance of model fits to these datasets. We will use two case studies to demonstrate the practicality and usefulness of supervised heatmaps in achieving these goals. The first will examine crime in US communities for which we will use the supervised heatmaps to gain an in-depth understanding of the information contained within the data, the clarity of which is unparalleled by existing visualization methods. The second case study will explore neural activity in the visual cortex where we will use supervised heatmaps to guide an exploration of the suitability of a Lasso-based linear model in predicting brain activity. Supervised heatmaps are implemented via the *superheat* package written in the R programming software.

*Keywords:* Data Visualization, Exploratory Data Analysis, Heatmap, Model Assessment, Multivariate Data

# 1 Introduction

The rapid technological advancements of the last few decades have enabled us to collect vast amounts of data. To accommodate these large datasets, we have begun to employ sophisticated computational methods in order to find answers to complex questions both in science and beyond. Although visualization has the capacity to be a powerful tool in the information extraction process of large multivariate datasets, there currently exist few tools capable of effectively representing high-dimensional datasets in a graphical domain. The majority of commonly used graphical exploratory techniques such as the traditional scatterplots, boxplots and histograms are embedded in spaces of 2 dimensions, and typically do not extend well to higher dimensions. Basic extensions into 3 dimensions are not uncommon, but are often inadequately represented when displayed in a 2-dimensional document (see Figure 9 for an example of a 3-dimensional plot displayed in 2 dimensions). New approaches to the visualization of high-dimensional data often subsequently face the difficult trade-off between encompassing high levels of complexity which may impede comprehension, and oversimplification leading to a loss of important information.

## *Existing multivariate data visualization approaches*

There have been a number of innovations within the statistics community in the field of visualizing and interpreting multivariate data (see Liu (2011); Chan (2006); Wong and Bergeron (1994) for reviews). At present, the most widely used multivariate data visualization techniques in statistics are the scatterplot matrix (Cleveland, 1993; Andrews, 1972) and parallel coordinates (Inselberg, 1985, 1998; Inselberg and Dimsdale, 1987).

Scatterplot matrices allow for multiple pairwise comparisons but do not allow for comparison of more than two variables at once for a single data point. Parallel coordinates plots, on the other hand, are capable of comparing multiple variables simultaneously over individual data points. However, both approaches become incomprehensible in the face of large numbers of observations as a result of over-plotting (an issue that can be mildly relieved with the use of transparency). For example, Figures 1 and 2 display a scatterplot matrix and parallel coordinates plot for 7 of the 32 variables from the *Communities and Crime* dataset (to be introduced in Section 3) which has approximately 2,000 observations. Even for this (not too large) dataset, using these existing tools it is difficult to effectively communicate the information and structure contained in the data.

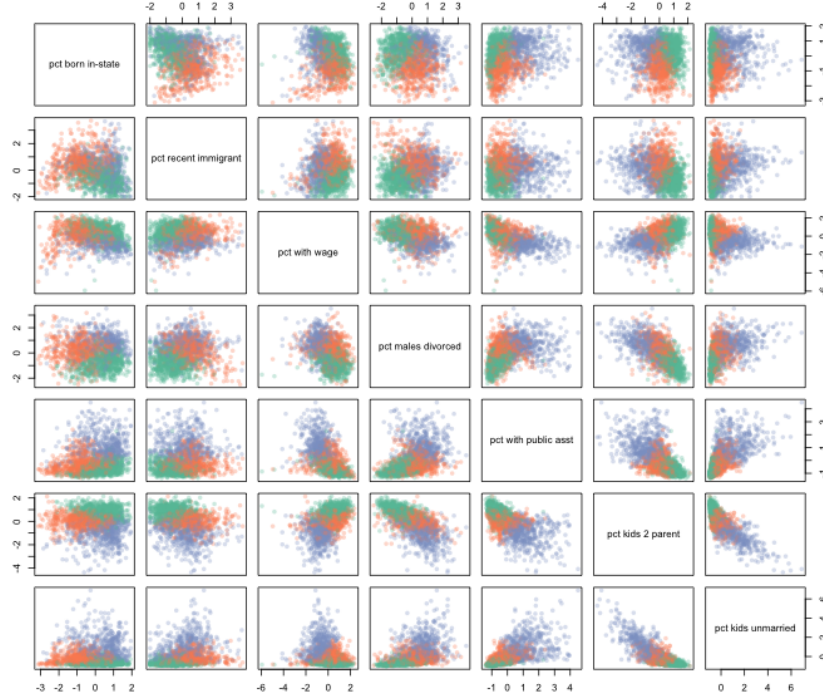


Figure 1: A scatterplot matrix for 7 of the 32 predictor variables from the *Communities and Crime* dataset which will be introduced in Section 3. The observations (corresponding to each point) are transparent and colored by clusters identified using a K-means clustering algorithm.

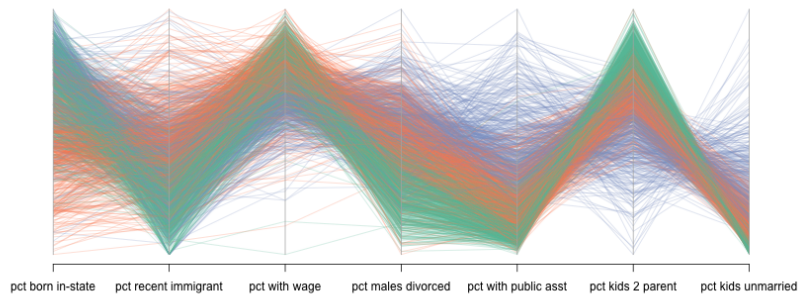


Figure 2: A parallel coordinates plot for 7 of the 32 predictor variables from the *Communities and Crime* dataset to be introduced in Section 3. The observations (corresponding to each connected line) are transparent and colored by clusters identified using K-means.

## *The heatmap*

Heatmaps are a high-dimensional data visualization tool widely used in areas such as bioinformatics (for example, to visualize large gene expression datasets), yet significantly underused in other areas of statistics (Wilkinson and Friendly, 2009). Heatmaps can be used to visualize matrices (such as the design matrix,  $X$ , whose columns correspond to predictor variables and whose rows correspond to observations) by representing each entry in the matrix by a color corresponding to its magnitude. Heatmaps enable the user to visually process large datasets with thousands of observations and/or variables without suffering from the problem of over-plotting that plagues existing methods such as parallel coordinates and scatterplot matrices. For larger matrices, it is common to use clustering as a tool to group together similar observations for the purpose of revealing structure in the data and increasing the clarity of information provided by the visualization (Eisen et al., 1998).

In this paper we introduce the supervised heatmap which builds upon the existing clustered heatmap widely used in bioinformatics. The primary contribution of our customizable supervised heatmaps is an extension to datasets which also have a response variable,  $y$ , such as in supervised learning or regression situations. In essence, supervised heatmaps have two primary uses: exploration of the structure of a dataset with reference to another variable such as a response variable, and the examination of the adequacy of the fit of data models.

### **1.1 Outline of the remainder of the paper**

The remaining sections are organized as follows: Section 2 introduces the supervised heatmap and describes its two primary uses. Sections 3 and 4 present two case studies that highlight the capabilities of the supervised heatmaps. The first case study in Section 3 comes from the *Communities and Crime* dataset created by Michael Redmond which can be found at the UCI machine learning repository. This dataset contains observations from approximately 2,000 communities in the US with over 30 predictor variables for violent crime rate. In Section 4, we examine functional magnetic resonance imaging (fMRI) data obtained from the Gallant Lab at UC Berkeley whereby we predict the brain’s response to viewings of 1,750 images each of which contain over 10,000 variables. Section 5 discusses how to implement and customize the supervised heatmaps in R using the R package `superheat` written by the authors. Section 6 concludes.

## 2 The supervised heatmap

The supervised heatmap proposed in this paper extends the existing clustered heatmaps widely used in bioinformatics. While these existing heatmap approaches offer a visualization of the design matrix,  $X$ , only, supervised heatmaps offer an extension to more general datasets that are accompanied (or are “supervised”) by a response variable,  $y$ , as in classification and regression problems.

Supervised heatmaps can be used to explore supervised data with two distinct purposes. The first is to examine the structure and relationships inherent within  $X$  and  $y$  together, and the second is to conduct model checking for a model,  $\hat{y} = f(X)$ , by examining the residuals,  $e = y - \hat{y}$ , within each cluster of the data (in general, one hopes to see that the residuals are not too large and are randomly scattered about the line  $e = 0$ ). The first application provides an in-depth understanding of the information contained in the data, while the second provides clues as to which regions of the data might be causing a model to perform sub-optimally or to suggest alternative modeling regimes.

Given a design matrix,  $X$ , and a response variable,  $y$ , the *supervised heatmap* consists of the following main features:

- a central heatmap, typically of the design matrix,  $X$ , or correlation matrix,  $cor(X)$ . The rows and/or columns of this heatmap are organized into clusters (either automatically or manually) to aid interpretability.
- a plot of the response,  $y$ , or of the residuals,  $e$ , located above or to the right of the heatmap (depending on whether it is the rows or the columns that correspond to the observations). These plots could be scatterplots, barplots, smoothed curves or lines.

### *Superheat: an R package*

It can be surprisingly difficult to manually combine and organize graphics such as those that appear in the supervised heatmap. To combat this limitation we have produced an R software package, **superheat**, which, on most laptop computers, can be used to comfortably produce highly customizable supervised heatmaps for datasets with several thousand rows and columns (the size constraints for plotting result from the pixel constraints of most computer screens). The

implementation of the `superheat` package will be discussed in Section 5. The following sections present two case studies which highlight the practicality and usefulness of the supervised heatmaps in (1) extracting and understanding information contained in high-dimensional datasets, and (2), examining the adequacy of the fit of data models.

### 3 Case study I: communities and crime

The *Communities and Crime* dataset created by Michael Redmond is hosted on the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>). The dataset contains socio-economic data for communities in the US obtained from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. There are a total of 2215 communities with over 100 predictor variables. Our response variable of interest is violent crimes per 100,000 population.

Within the set of predictor variables, there turned out to be several large groups of variables each conveying the same information (such as the percentages of people born in their current state, people living in the same house as in 1985 and people living in the same state as in 1985). After randomly selecting one variable from each group of similar variables (where two variables are deemed to be “similar” if they have a correlation above 0.7), the dataset has a total of 32 predictor variables (such as the percentage of households with wages and the percentage of recent immigrants; see Appendix B for a full list) and a single response variable (violent crime rate per 100,000 population). Figures 1 and 2 present a subset of this dataset using the existing scatterplot matrix and parallel coordinate plot approaches.

#### 3.1 Exploratory data analysis using supervised heatmaps

Figure 3 displays a supervised heatmap based on the standardized community variables. Heatmaps cannot effectively represent variables that are on different scales, so each variable is standardized to have mean 0 and variance 1. The heatmap shows the relationship between each predictor variable as well as to the response variable (violent crimes per 100K population) represented by a loess-smoothed curve and scatterplot above the heatmap.

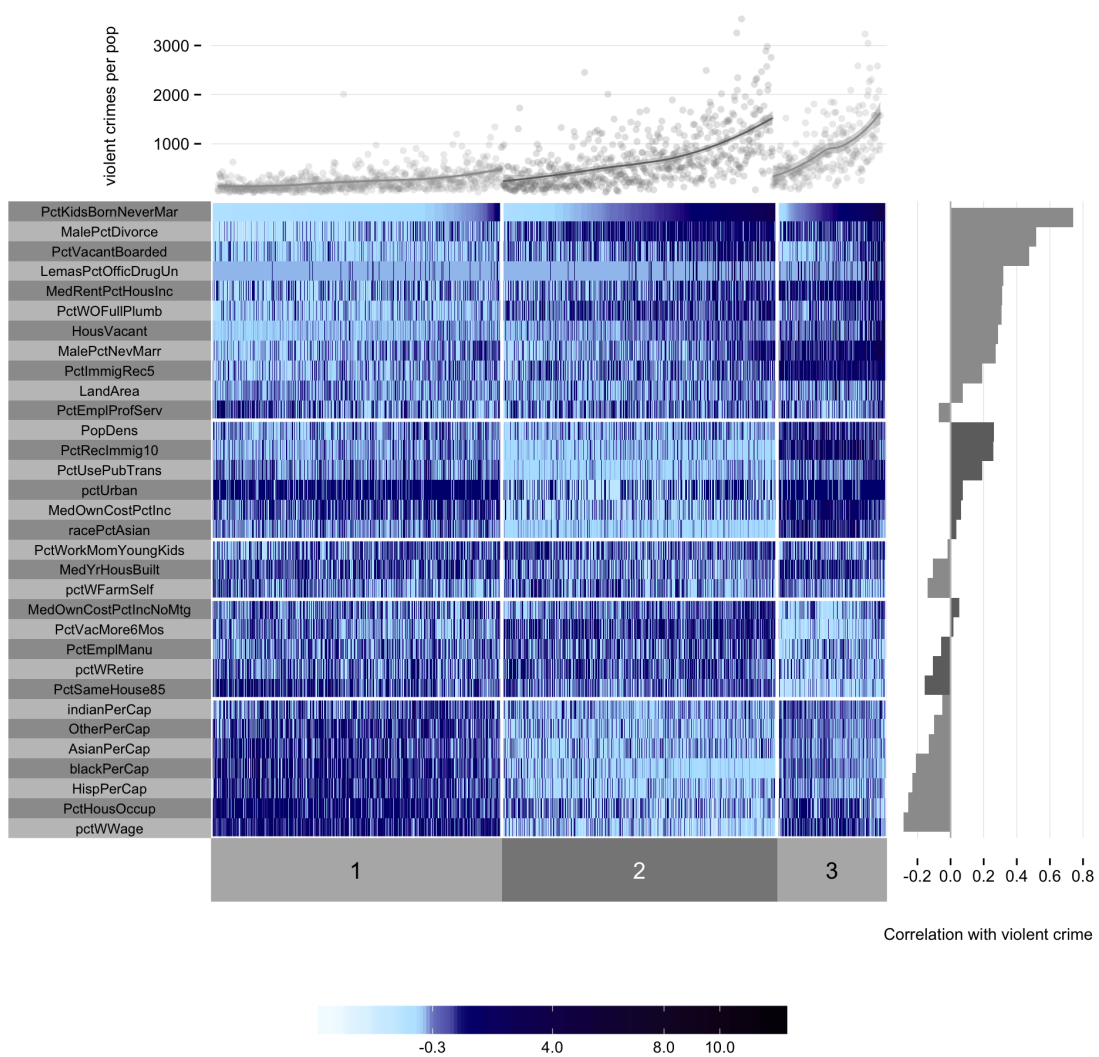


Figure 3: A supervised heatmap for the Communities and Crime dataset. The heatmap corresponds to the matrix whose rows correspond to the (standardized) predictor variables grouped into five clusters and whose columns are the communities grouped into three clusters. Above the heatmap is a scatterplot overlaid with a loess curve of the violent crime rate within each community where the  $x$ -axis corresponds to the percentage of households in the community with children born to unmarried parents. On the right-hand side, a barplot shows the correlation of each variable with the violent crime rate.

### *Examination of variable clusters*

The rows (predictor variables) of the heatmap are arranged into five clusters of communities generated using *K*-means (the default clustering algorithm implemented in the `superheat` package), and the correlation of each variable with the response (violent crime rate) is presented as a bar plot to the right of the heatmap. The first cluster of variables (containing the percentages of children born to unmarried parents, divorced men, vacant houses, etc) are mostly positively correlated with the response variable, whereas the fifth cluster of variables (consisting of several variables related to race as well as percentage of houses occupied) are slightly negatively correlated with violent crime rate.

### *Examination of community clusters*

The columns (observations/communities) of the heatmap are arranged into three clusters of communities. Within each cluster, the order of the columns/communities (as well as the *x*-axis of the scatterplot above the heatmap) corresponds to the proportion of children born to unmarried parents in the community (the first variable in the heatmap). From Figure 3, the first cluster of communities appear to mostly be urban communities with a relatively low rate of divorce, few vacant houses, few recent immigrants and a high proportion of households earning wages. This cluster of communities enjoys a comparatively low violent crime rate.

Compared to the first cluster of communities, the communities in the second cluster tend to be less urban, less racially diverse, tend to have higher values for the variables in the first cluster (e.g. percentage of divorce and children born to unmarried parents), but have lower values of almost all variables in the second and fifth clusters (e.g. percentage of recent immigrants and occupied houses). The crime rate for these communities is higher than those in the first cluster and increases with the proportion of children born to unmarried parents. Note that we do not claim that such variables *cause* an increase in violent crime, only that the two are correlated.

The third cluster of communities (which includes larger urban cities such as New York City, San Francisco and Los Angeles), on the other hand, has higher rent as a percentage of household income, a higher percentage of recent immigrants and unmarried males as well as fewer vacant houses and fewer retirees than the other two community clusters. This cluster of communities experiences an increase in crime rate similar to that of the second cluster of communities.



### 3.2 Assessment of model fit using supervised heatmaps

We next decided to fit two linear models both of which could be used to predict violent crime rate in new communities. The first linear model uses the untransformed violent crime rate variable as the response and the second linear model uses a log-transformed violent crime rate,  $\log(1 + \text{ViolentCrimesPerPop})$ , as the response. To perform feature selection, we fit a Lasso linear model to each response, using all 32 of the variables (where the variables were each scaled to have mean 0 and variance 1). In the untransformed model, six of the variable coefficients were shrunk to zero by the Lasso estimator, while in the log-transformed model, nine of the variable coefficients were shrunk to zero. Thus our linear models contained 26 and 23 predictor variables, respectively.

Figure 4 presents a supervised heatmap which we will use to identify which model is a better fit for the data. Instead of plotting a heatmap of the design matrix as in the previous example (where the rows corresponded to the variables and the columns corresponded to the observations/communities), in this example, we plot a heatmap of the correlation matrix (both the rows and the columns correspond to the communities organized into the same three clusters that were present in Figure 3). In particular, the color of cell  $(i,j)$  in the heatmap in Figure 4 corresponds to the correlation between communities  $i$  and  $j$ .

Figure 4 shows that the communities within each cluster are fairly correlated with one another and are mostly negatively correlated with communities in the other clusters, reflecting the effectiveness of the default K-means clustering algorithm.

Next, we plot the residuals for the untransformed linear model for violent crime rate above the heatmap and we plot the residuals for the log-transformed linear model to the right. The residuals in each plot (as well as the columns and rows of the heatmap) are sorted within each cluster by the fitted values for the corresponding linear model.

Evaluation of the top residual plot in Figure 4 reveals that the standard (untransformed) linear model does not fit the data well. In particular, the residuals for the first cluster have a downward linear trend and the residuals for the second and third clusters exhibit heteroskedasticity. On the other hand, the log-transformed linear model appears to be a much better fit for the data (see the right residual plot in Figure 4): the residuals appear to be randomly scattered about zero and exhibit constant variance.

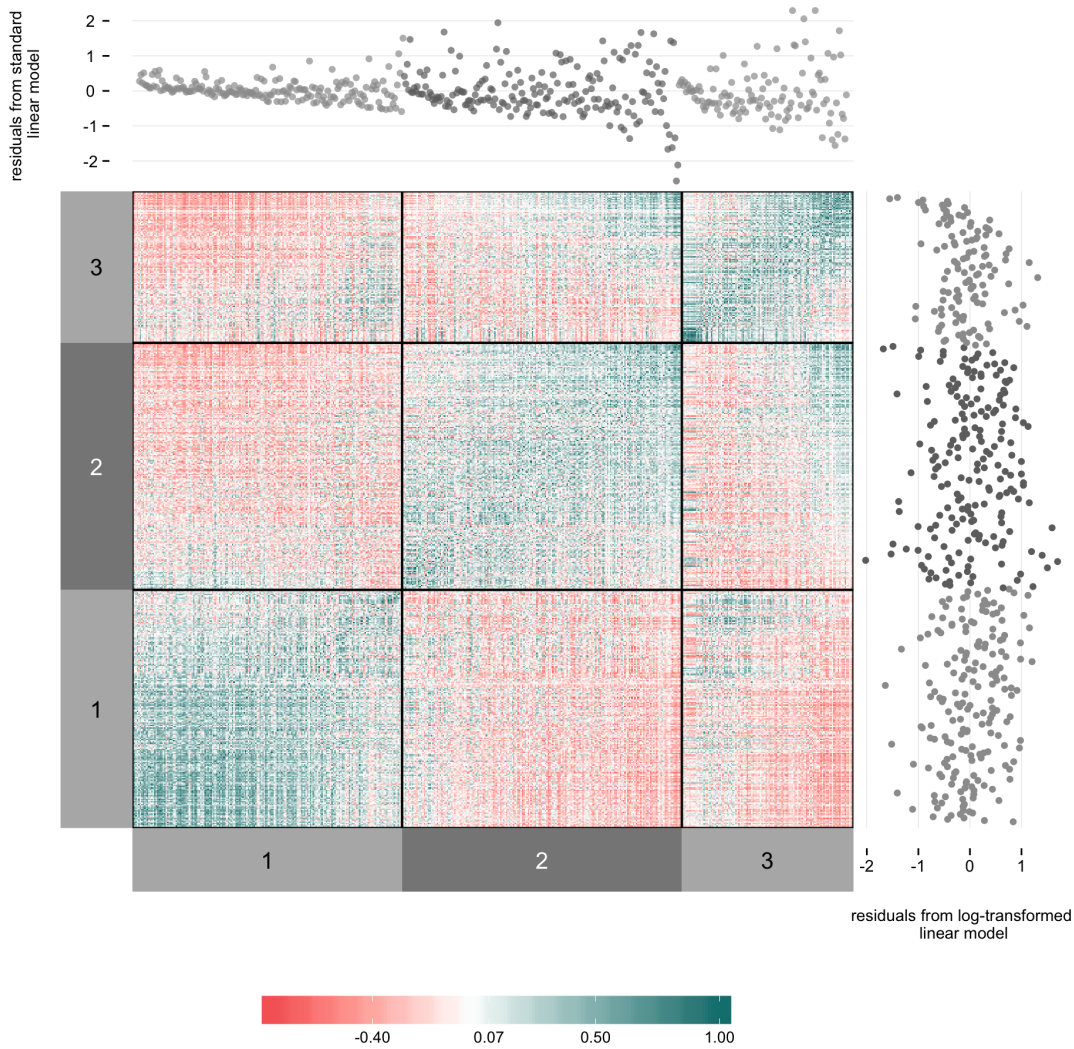


Figure 4: A supervised heatmap comparing a standard linear model and a log-transformed linear model for violent crime rate from the Communities and Crime dataset. The heatmap corresponds to a correlation matrix whose rows and columns correspond to the observations/communities in our datasets. That is, entry  $(i,j)$  in the heatmap corresponds to the correlation between communities  $i$  and  $j$ . Both the rows and columns have been sorted into the clusters described in Figure 3. Above the heatmap is a residual plot for a standard linear model for violent crime rate. To the right of the heatmap is an equivalent residual plot for the log-transformed linear model. The axes are sorted by the fitted values of the corresponding model within each cluster.

## 4 Case study II: neuroscience

Our second case study examines data collected from a functional magnetic resonance imaging (fMRI) experiment performed on a single individual by the Gallant neuroscience lab at UC Berkeley (Vu et al., 2009, 2011). fMRI measures oxygenated bloodflow in the brain which can be considered as an indirect measure of neural activity (the two processes are highly correlated). The measurements obtained from an fMRI experiment correspond to the aggregated response of hundreds of thousands of neurons within cube-like voxels of the brain, where the segmentation of the brain into voxels is analogous to the segmentation of an image into pixels.

The data contains the fMRI measurements (averaged over 10 runs of the experiment) for each of 20 voxels located in the visual cortex of a single individual in response to viewings of 1750 different images (such as a picture of a baby, a house or a horse). Each image is a  $128 \times 128$  pixel grayscale image which can be represented by a vector of length  $128^2 = 16384$  but can be reduced to length 10921 through a Gabor wavelet transformation (Lee, 1996). See Figure 5 for a graphical representation of the data structure.

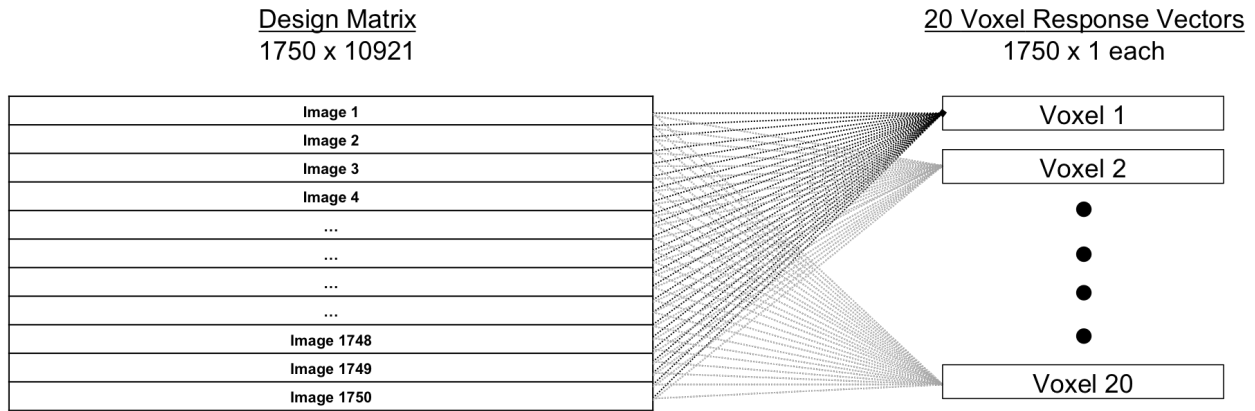


Figure 5: This diagram describes the fMRI data: we have a design matrix with 1750 observations (images) and 10921 features (Gabor wavelets) for each image. We also have 20 distinct response vectors (voxel responses) where for each voxel we have recorded the average (indirect brain activity) response to each of the 1750 images. Thus we fit 20 different models, one for each voxel, where each model has the same design matrix. The heatmap in Figure 6 corresponds to the design matrix whereas the heatmap in Figure 8 corresponds to the voxel response matrix.

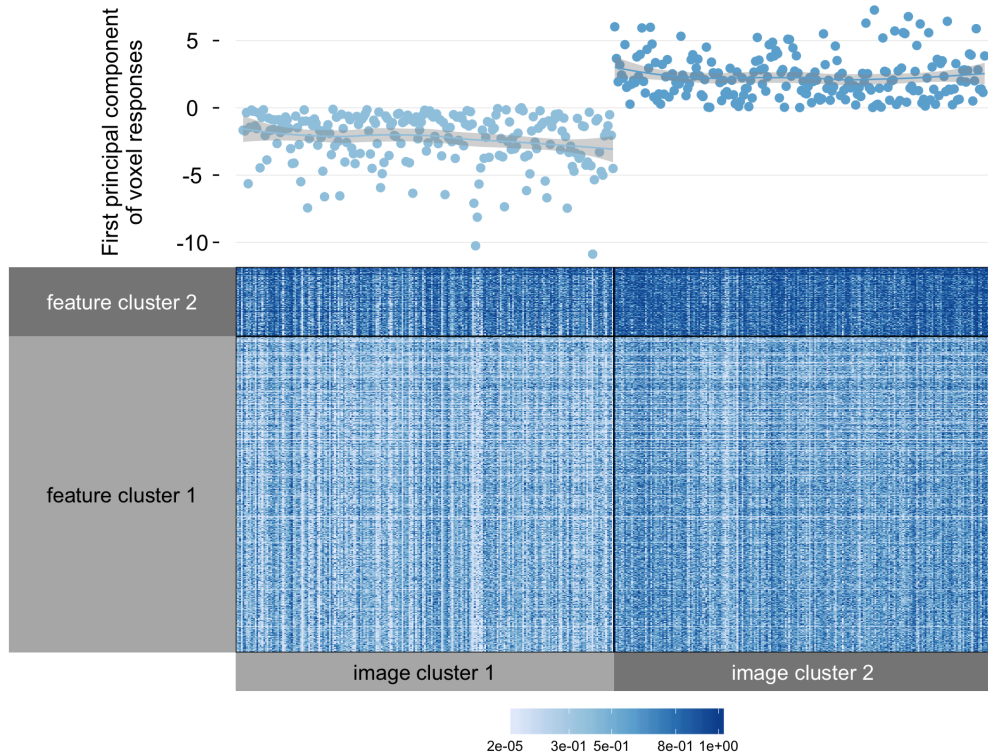


Figure 6: A supervised heatmap for the fMRI data. The feature values are plotted in the heatmap where the columns correspond to the 1750 images and the rows correspond to a randomly chosen subset of 500 of the 10921 Gabor wavelet features. Both the columns and the rows are grouped into two clusters. The first principal component of the 20 voxel responses for each image is presented in a scatterplot above the heatmap.

#### 4.1 Exploratory data analysis using supervised heatmaps

Figure 6 presents a supervised heatmap for which the central heatmap corresponds to a randomly chosen  $500 \times 1750$  subset of the  $10921 \times 1750$  design matrix. That is, the columns correspond to the images and the rows correspond to a random sample of 500 of the the Gabor wavelet features. We consider a subset of the data since most laptop computers have a screen resolution consisting of just over one million pixels, and as such the supervised heatmap can comfortably handle matrices with up to one million entries. Both the features and images are segmented into two distinct clusters generated using K-means (the default clustering algorithm of the `superheat` package).

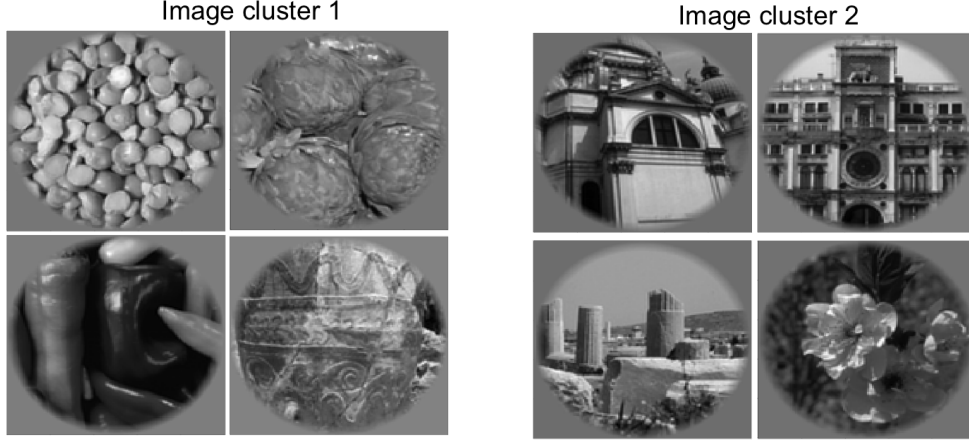


Figure 7: Four randomly selected images from each cluster in the fMRI experiment.

The second cluster of images shown in 6 exhibit higher Gabor wavelet feature values than the images in the first cluster. Figure 7 displays four randomly selected examples of images from each cluster. In these examples (as with other randomly chosen samples), the images in the first cluster tend to exhibit less variation within the image and the contents of the images tend to be less recognizable than the images in the second cluster. This observation reflects the fact that Gabor wavelets capture salient visual properties such as discontinuity in gradient (Wei and Bartels, 2006).

In the scatterplot above the heatmap in Figure 6, for each image (column) we have plotted the linear combination of the 20 voxel responses corresponding to the first principal component (the first principal component accounts for approximately 43% of the total variability in the data; see Appendix C for a scree plot of the first 10 principal components).

Interestingly, the voxel responses to the images in the second cluster are much larger than the responses to images from the first cluster. That is, the brain is more active when viewing images from the second cluster which also had higher Gabor wavelet feature values than the images in the first cluster.

## 4.2 Assessment of model fit using supervised heatmaps

Next, since the data contains 10,921 features, in the interests conducting feature selection we fit Lasso linear models to each of the 20 voxels using the Gabor wavelet features as the predictor variables (See Figure 5). The Lasso will shrink the majority of the model coefficients to zero.

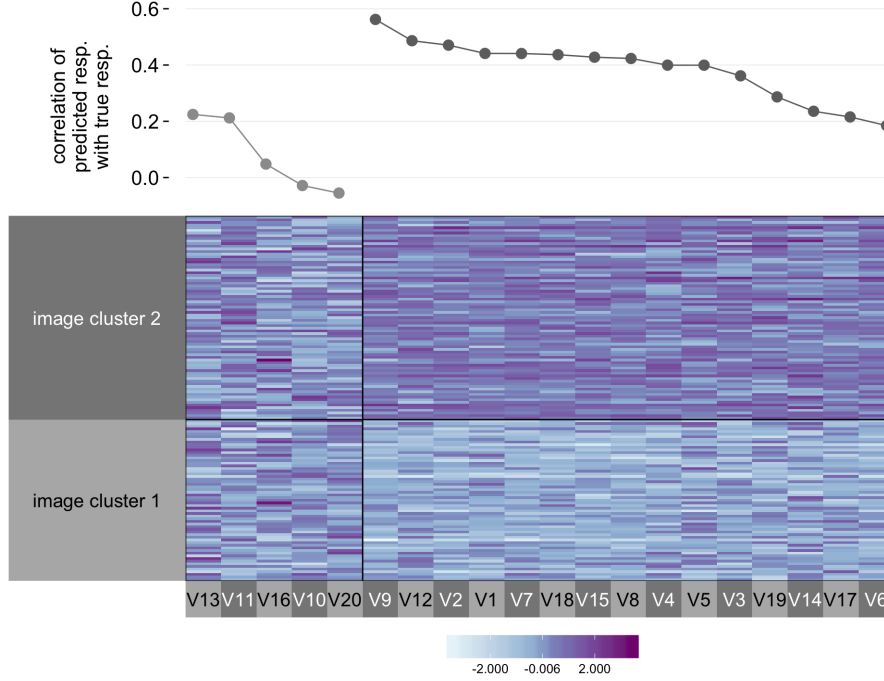


Figure 8: A supervised heatmap examining the fit of the Lasso linear model for each voxel. The heatmap displays the response of each voxel (column) to each image (row). The correlation of the predicted response (based on a withheld test set) with the observed response is plotted above.

The regularization parameters for the Lasso were selected using an estimation stability metric combined with cross-validation (ESCV) (Lim and Yu, 2015). The number of Gabor features selected by these Lasso models (the number of Gabor features with nonzero coefficients) ranged from 5 for voxel 19 to 183 for voxel 7. The heatmap in Figure 8 displays the voxel responses to each image (this information was presented in an aggregated form in the scatterplot above the heatmap in Figure 6). The columns of the heatmap correspond to the voxels, and so each column can be thought of as the response vector for a single model. The voxels are split into two clusters (using the default K-means algorithm), and the images are separated into the same clusters that appeared in Figure 6.

The first cluster of voxels do not appear to respond differently to the two clusters of images, whereas the second cluster of voxels are more active in response to the second cluster of images. In particular, the differences in voxel responses between the two clusters of images from Figure 6 appears to be driven solely by the voxels in the second cluster.

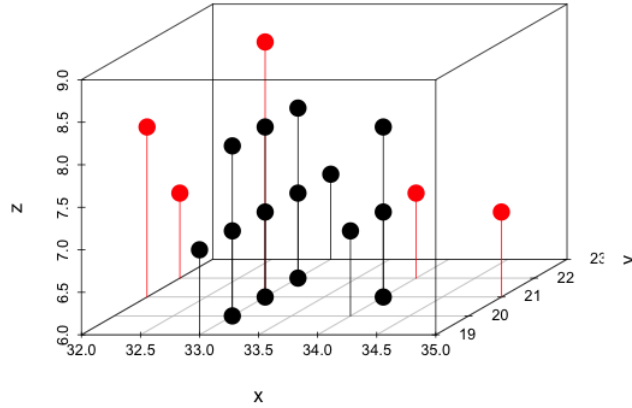


Figure 9: A 3D plot displaying the locations of the 20 voxels in the visual cortex. The red points correspond to the voxels in the first cluster, whereas the black points correspond to the voxels in the second cluster. Each point is connected to the x-y plane by a vertical line.

Above the heatmap, the correlation of the predicted response (based on a withheld test set) with the actual observed average response for each voxel is plotted in decreasing order. It is clear that the responses to the first cluster of voxels (voxels 10, 11, 13, 16 and 20) are not adequately predicted by the Lasso linear models, whereas the Lasso models for the voxels in the second cluster appear to perform much better. This finding suggests that the voxel responses in the second cluster are driven more by the Gabor features of the images than the voxels in the first cluster. Prompted by these findings, further investigation revealed that the voxels in the first cluster were actually located on the very edge of the visual cortex (see Figure 9), and thus may contain many neurons that are outside of the visual cortex and are thus less relevant for neural image processing.

## 5 Implementation of supervised heatmaps

As evident from the examples presented above, supervised heatmaps are flexible, customizable and very useful. Such plots would be difficult and time-consuming to produce without the existence of software that can automatically generate the plots given the user’s preferences. **superheat** is a software package written by the authors that implements supervised heatmaps in the R programming language. The package makes use of the popular **ggplot2** package, but does not utilize the **ggplot2** grammar of graphics (Wickham, 2010). In particular, **superheat** exists as a stand-alone



function with the sole purpose of producing customizable supervised heatmaps. The development page for **superheat** is hosted openly at <https://github.com/rlbarter/superheat>, where the user can also find a detailed Vignette describing further information on the specific usage of the plot in R as well as a host of options for customizability. See Appendix A for some example code.

There are two main types of customizability for the **superheat** package: cluster implementation and aesthetics. To customize the cluster implementation, the user has the option to use their own clustering algorithm by providing a predefined membership vector. If the user does not wish to perform their own clustering, the default is to cluster the rows using K-means on the correlation matrix (the columns are not clustered unless otherwise specified). To select the number of clusters, it is recommended that the user does so prior to the implementation of the supervised heatmaps using standard methods such as Silhouette plots (Rousseeuw, 1987).

A vast number of aesthetic options exist for the supervised heatmaps. For instance, each of the figures presented in this paper exhibited unique color schemes (this is possible, not only for the color scale in the heatmap, but also for the plots above and to the right of the heatmap as well as the text and labels for the plots). Moreover, there are several options for the form of the top and right plots: scatterplots (the default), scatterplots with a smoothed curve, an isolated smoothed curve, barplots, line plots and scatterplots with points connected by lines. These options, and more, are demonstrated in the Vignette that can be found on the github page.

## 6 Conclusion

Heatmaps are an extremely useful data visualization tool, particularly for high-dimensional datasets. Supervised heatmaps augment traditional heatmaps via the inclusion of extra information, such as a response variable or residuals, providing the user with an additional instrument for information extraction. Supervised heatmaps, as implemented by the **superheat** package written by the authors, are highly customizable and can be used effectively in a wide range of situations such as exploratory data analysis and model assessment. The usefulness of the supervised heatmaps was highlighted in two case studies, the first assessing the relationship between various socio-economic variables and violent crime rate in US communities and the second examined fMRI data capturing the brain's response to viewings of images.



## A APPENDIX: IMPLEMENTATION IN R

The `superheat` package introduces a new function (conveniently also named `superheat`). The use of this function to generate Figure 3 in this paper is shown below. The code to generate the remaining supervised heatmap figures can be found in the supplementary materials.

# Figure 3: A supervised heatmap of the communities and crime dataset.

```
superheat(X = t(crime),
          n.clusters.cols = 3,
          heat.pal = c("white", "lightskyblue1",
                       "navyblue", "black"),
          heat.pal.values = c(0, 0.4, 0.6, 1),
          order.cols = order(t(crime)["pct kids unmarried",]),
          cluster.rows = FALSE,
          box.col = "white",
  # left labels
          left.text.angle = 0,
          left.label.size = 0.4,
  # top plot
          yt = violent_crimes,
          yt.num.ticks = 4,
          yt.plot.type = "scattersmooth",
          yt.axis.name = "violent crimes per pop",
          yt.point.alpha = 0.2,
  # right plot
          yr = cor.vc,
          yr.plot.type = "bar",
          yr.axis.name = "Correlation with violent crime",
          yr.num.ticks = 4,
  # legend
          legend.size = 4)
```

## B APPENDIX: VARIABLES FROM THE COMMUNITIES AND CRIME DATASET

|    | Variable name         | Description   |
|----|-----------------------|---|
| 1  | AsianPerCap           | per capita income for people with Asian heritage  |
| 2  | blackPerCap           | per capita income for African Americans   |
| 3  | HispPerCap            | per capita income for people with Hispanic heritage                                       |
| 4  | HousVacant            | number of vacant household  |
| 5  | indianPerCap          | per capita income for native Americans  |
| 6  | LandArea              | land area in square miles   |
| 7  | LemasPctOfficDrugUn   | percent of officers assigned to drug units  |
| 8  | MalePctDivorce        | percentage of males who are divorced  |
| 9  | MalePctNevMarr        | percentage of males who have never married  |
| 10 | MedOwnCostPctInc      | median owners cost as a percentage of household income<br>- for owners with a mortgage    |
| 11 | MedOwnCostPctIncNoMtg | median owners cost as a percentage of household income<br>- for owners without a mortgage |
| 12 | MedRentPctHousInc     | median gross rent as a percentage of household income                                     |
| 13 | MedYrHousBuilt        | median year housing units built   |
| 14 | OtherPerCap           | per capita income for people with ‘other’ heritage  |
| 15 | PctEmplManu           | percentage of people 16 and over who are employed in<br>manufacturing                     |
| 16 | PctEmplProfServ       | percentage of people 16 and over who are employed in<br>professional services             |
| 17 | PctHousOccup          | percent of housing occupied   |
| 18 | PctImmigRec5          | percentage of immigrants who immigrated within last 5<br>years                            |
| 19 | PctKidsBornNeverMar   | percentage of kids born to never married  |

|    | Variable name       | Description  |
|----|---------------------|--|
| 20 | PctRecImmig10       | percent of population who have immigrated within the last 10 years   |
| 21 | PctSameHouse85      | percent of people living in the same house as in 1985                |
| 22 | pctUrban            | percentage of people living in areas classified as urban             |
| 23 | PctUsePubTrans      | percent of people using public transit for commuting                 |
| 24 | PctVacantBoarded    | percent of vacant housing that is boarded up                         |
| 25 | PctVacMore6Mos      | percent of vacant housing that has been vacant more than 6 months    |
| 26 | pctWFarmSelf        | percentage of households with farm or self employment income in 1989 |
| 27 | PctWOFullPlumb      | percent of housing without complete plumbing facilities              |
| 28 | PctWorkMomYoungKids | percentage of moms of kids 6 and under in labor force                |
| 29 | pctWRetire          | percentage of households with retirement income in 1989              |
| 30 | pctWWage            | percentage of households with wage or salary income in 1989          |
| 31 | PopDens             | population density in persons per square mile                        |
| 32 | racePctAsian        | percentage of population that is of Asian heritage                   |

Table 1: A list of the 32 variable names and definitions for the Communities and Crimes dataset.

## C APPENDIX: SCREE PLOT FOR THE PCA OF VOXEL RESPONSES

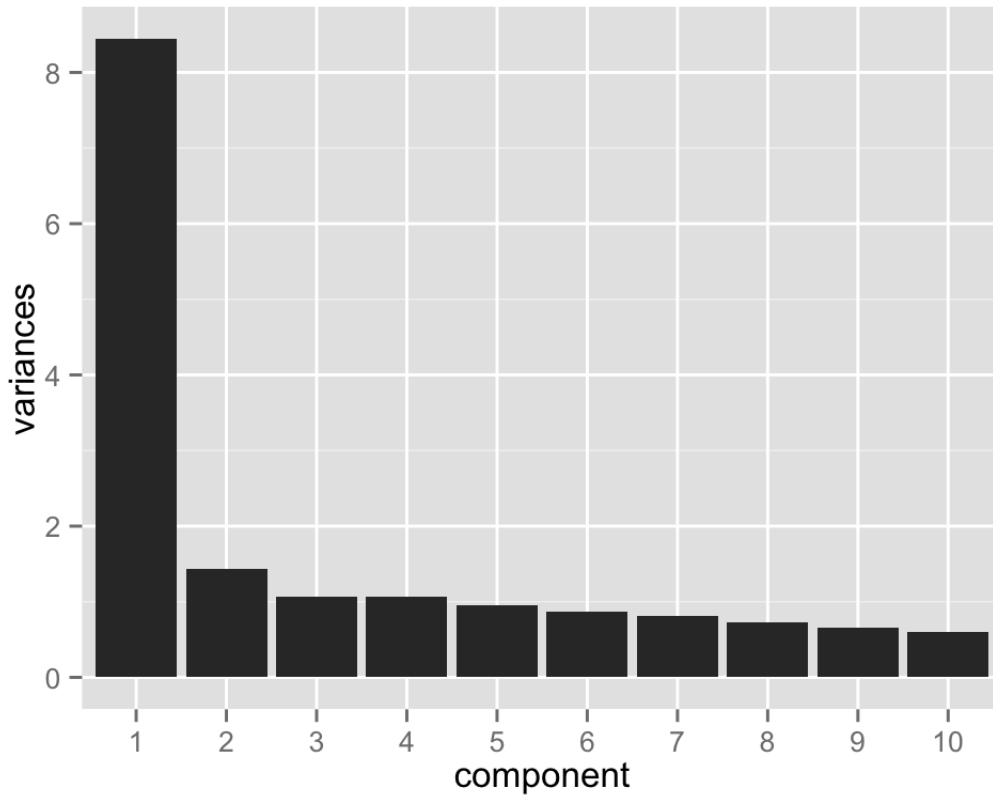


Figure 10: A scree plot displaying the decay in eigenvalues/variance for each of the first 10 principal components for the voxel responses.

## SUPPLEMENTARY MATERIAL

**R code:** 0.clean.R, 1-figures.R, R scripts that load and clean the *Communities and Crime* and fMRI data and make the figures that appear in this paper. (R scripts)

## ACKNOWLEDGMENTS

The authors would like to thank the Gallant Lab at UC Berkeley for providing the fMRI. This research is partially supported by NSF grants DMS-1107000, CDS&E-MSS 1228246, DMS-1160319 (FRG), NHGRI grant 1U01HG007031-01 (ENCODE), AFOSR grant FA9550-14-1-0016, and the Center for Science of Information (CSoI), an USNSF Science and Technology Center, under grant agreement CCF-0939370.

## References

- Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics* 28(1), 125–136.
- Chan, W. W.-Y. (2006). A survey on multivariate data visualization. *Department of Computer Science and Engineering. Hong Kong University of Science and Technology* 8(6), 1–29.
- Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.
- Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95(25), 14863–14868.
- Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer* 1(2), 69–91.
- Inselberg, A. (1998). Visual data mining with parallel coordinates. *Computational Statistics* 13(1).
- Inselberg, A. and B. Dimsdale (1987). Parallel coordinates for visualizing multi-dimensional geometry. In D. T. L. Kunii (Ed.), *Computer Graphics 1987*, pp. 25–44. Springer Japan.
- Lee, T. S. (1996). Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 18(10), 959–971.

- Lim, C. and B. Yu (2015). Estimation stability with cross validation (escv). *Journal of Computational and Graphical Statistics* 0(just-accepted).
- Liu, Y. (2011). Multivariate data visualization: a review from the perception aspect. In *Human Interface and the Management of Information. Interacting with Information*, pp. 221–230. Springer.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.
- Vu, V. Q., P. Ravikumar, T. Naselaris, K. N. Kay, J. L. Gallant, and B. Yu (2011). Encoding and decoding v1 fmri responses to natural images with sparse nonparametric models. *The Annals of Applied Statistics* 5(2B), 1159–1182.
- Vu, V. Q., B. Yu, T. Naselaris, K. Kay, J. Gallant, and P. K. Ravikumar (2009). Nonparametric sparse hierarchical models describe v1 fmri responses to natural images. *Advances in Neural Information Processing Systems (NIPS)* 21, 1337–1344.
- Wei, H. and M. Bartels (2006). Unsupervised segmentation using gabor wavelets and statistical features in lidar data analysis. In *18th International Conference on Pattern Recognition, 2006. ICPR 2006*, Volume 1, pp. 667–670.
- Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics - J COMPUT GRAPH STAT* 19(1).
- Wilkinson, L. and M. Friendly (2009). The history of the cluster heat map. *The American Statistician* 63(2), 179–184.
- Wong, P. C. and R. D. Bergeron (1994). 30 years of multidimensional multivariate visualization. In *Scientific Visualization*, pp. 3–33.